On the effect of fluctuating recombination rates on the decorrelation of gene histories in the human genome

A. Eriksson and B. Mehlig

Physics and Engineering Physics, Gothenburg University/Chalmers, 41296 Gothenburg, Sweden

Abstract

We show how to incorporate fluctuations of the recombination rate along the chromosome into standard gene-genealogical models for the decorrelation of gene histories. This enables us to determine how small-scale fluctuations (Poissonian hot-spot model) and large-scale variations (KONG *et al.*, 2002) of the recombination rate influence this decorrelation. We find that the empirically determined large-scale variations of the recombination rate give rise to a significantly slower decay of correlations compared to the standard, unstructured gene-genealogical model assuming constant recombination rate. A model with long-range recombination-rate variations and with demographic structure (divergent population) is found to be consistent with the empirically observed slow decorrelation of gene histories. Conversely, we show that small-scale recombination-rate fluctuations do not alter the large-scale decorrelation of gene histories.

Genome-wide variation and decorrelation of gene histories are reflected in patterns of linkage disequilibrium which in turn shape the genetic variation observed on the molecular level. Recently REICH *et al.* (2002) reported on the first genome-wide measurement of correlations of human gene histories. REICH *et al.* (2002) show that their data are inconsistent with standard gene-genealogical models allowing for non-trivial population structures and demographic schemes, but assuming a constant recombination rate over the genome. The question is thus: can fluctuations of the recombination rate along the chromosome explain the slow correlation decay of gene histories?

Empirical results (CHAKRAVARTI *et al.*, 1984; GOLDSTEIN, 2001; JEFFREYS *et al.*, 2001) indicate that the recombination rate is not constant along the chromosome. It was observed that, at certain locations, an appreciable fraction of recombination events are concentrated in short regions [roughly 1kb long and spaced 50kb apart (STUMPF and GOLDSTEIN, 2003)], so-called hot spots. At least locally this implies

small-scale (< 100kb) variations of the recombination rate along the chromosome. Genome-wide, long-range fluctuations of the recombination rate for humans have been empirically determined by KONG *et al.* (2002). It is thus necessary to incorporate the effect of fluctuating recombination rates into the standard gene-genealogical model (GRIFFITHS, 1981; HUDSON, 1983; TAVARÉ, 1984; KAPLAN and HUDSON, 1985; HUDSON, 1990; NORDBORG and TAVARÉ, 2002). More generally, it is necessary to determine: on which length scales do recombination-rate fluctuations at a certain scale influence the decorrelation function of gene histories most significantly? It has been argued (REICH *et al.*, 2002) that small-scale recombination-rate fluctuations (< 100kb) related to hot spots are an important if not the main feature determining the slow decorrelation of gene histories (assuming that hot-spots are to be found genome-wide).

Here we derive an expression for the correlation of gene histories in neutral gene-genealogical models allowing for fluctuating recombination rates. This enables us to explain and quantitatively describe the influence of recombination-rate fluctuations on the correlation of gene histories. We find that large-scale fluctuations of empirically determined recombination rates (KONG *et al.*, 2002) give rise to a significantly slower decay of correlations compared to the standard, unstructured, constant population-size gene-genealogical model assuming constant recombination rate. Furthermore, a model with large-scale recombination-rate fluctuations and with demographic structure [divergent population, see (EYRE-WALKER *et al.*, 1998; TESHIMA and TAJIMA, 2002; REICH *et al.*, 2002) and references cited therein] is found to be consistent with the empirically observed decorrelation of gene histories. It is not necessary to invoke hot spots.

In a neutral model, KAPLAN and HUDSON (1985) [see also (GRIFFITHS, 1981)] have derived a relation between the correlation function ρ_{τ_x,τ_y} of the times τ_x and τ_y to the most recent common ancestors of two loci x and y, and the amount C of recombination between these two loci[1]. We observe that their result depends on the total amount of recombination between the two loci, but not on the distribution of recombination events between these loci. Moreover, this is still true when population structure is taken into account. The expected correlation $\rho^{\exp}_{\tau_x,\tau_y}$ is obtained by averaging with a sliding window of length $|y-x|$ along the chromosome. Thus, if $p_X(C)$ is the genome-wide distribution of recombination intensity C in bins of lengths $X = |y-x|$, the expected correlation is

$$\rho^{\exp}_{\tau_x,\tau_y} = \int \mathrm{d}C\, p_X(C)\, \rho_{\tau_x,\tau_y}(C)\,. \tag{1}$$

It also follows that small-scale fluctuations of the recombination rate on length scales much smaller than X

[1]The result of KAPLAN and HUDSON (1985) for the unstructured, constant population-size model is exact for sample size $n = 2$; for large n it is a very good approximation.

are irrelevant to the decay of correlations on scales of the order of X. In particular, fluctuations due to hot spots at small scales cannot change the decorrelation of gene histories at much larger scales.

Using (1) we have computed $\rho^{\rm exp}_{\tau_x,\tau_y}$ in four models (Fig. 1): assuming small-scale variation of the recombination rate (model I), incorporating, in addition, large-scale variation (model II), and estimating $p_X(C)$ from the empirical data of KONG *et al.* (2002) (model III), and, in addition, taking into account demographic population structure (model IV). Model I is the Poissonian hot-spot model of REICH *et al.* (2002), described in more detail in Fig. 1a below. From (1) we obtain an explicit expression for $\rho^{\rm exp}_{\tau_x,\tau_y}$ (caption of Fig. 1a). This result is shown in Fig. 2a. In agreement with REICH *et al.* (2002) we find that on distances of the order of the hot-spot spacing, correlations are larger than those in a constant recombination-rate model. However, there is no choice of parameters which could explain the empirically observed decorrelation function. (c.f. data in Fig. 6a of REICH *et al.* (2002), reproduced in Fig. 2b below). In particular, no significant increase in correlations on length scales $\gg \lambda^{-1}$ is observed, as discussed above.

REICH *et al.* (2002) have fitted an "arbitrary mixed model" to their empirical data. In order to obtain these results it is necessary to introduce large-scale variations of the recombination rate, on a scale $L \gg X \sim 1$Mb. One possibility (model II) is to assume that hot-spots occur in clusters, with long ($\gg 1$Mb) regions of low recombination intensity between them, see Fig. 1b. This model provides a better fit to the empirical data (Fig. 2b) than model I, indicating that large-scale fluctuations of the recombination rate are important. Notice that assuming $p_X(C) = (1-p)\,\delta(C - R_0 X) + p\,\delta(C - R_1 X)$ can produce an equally good fit to the data (e. g. for $p = 0.55$, $R_0 = 1.2$cM/Mb and $R_1 = 0.02$cM/Mb, not shown). In this model the recombination rate is constant on large scales ($\gg 1$Mb) and alternates between two values R_0 and R_1. However this model is not consistent with the empirically observed $p_X(C)$. We have estimated $p_X(C)$ from empirical data (KONG *et al.*, 2002) (Fig. 1c, model III). The corresponding results are shown in Fig. 2b. We find that the empirically determined large-scale fluctuations of the recombination rate give rise to significantly enhanced correlations (compared to the standard model assuming constant recombination rate), especially at large distances.

It is expected that population structure can increase the correlations of gene histories at large distances. We have considered the effect of large-scale recombination-rate fluctuations within a well-established model of demographic structure: the population was of constant size N until τ_0 generations ago, when it split into two fractions of size γN and $(1-\gamma)N$. The two sub-populations remained separate until a recent merging (see for instance EYRE-WALKER *et al.* (1998); TESHIMA and TAJIMA (2002); REICH *et al.*

(2002) and references therein). For sample size $n = 2$ we have calculated $\rho_{\tau_x,\tau_y}(C)$ explicitly in this model (ERIKSSON and MEHLIG, 2004). Without recombination-rate fluctuations, this model does not describe the empirically observed correlation of gene histories (REICH *et al.*, 2002). We have determined the effect of large-scale recombination-rate fluctuations (KONG *et al.*, 2002) on the correlation of gene histories in this model using eq. (1) and the explicit expression for $\rho_{\tau_x,\tau_y}(C)$. The parameters of the model (τ_0 and N) where chosen to be consistent with the empirically estimated time to the most recent common ancestor and its coefficient of variation (REICH *et al.*, 2002). The parameter γ was set to 0.3. The resulting correlation function matches the empirical data reasonably well. Decreasing γ gives rise to decreased correlations ($\gamma = 0$ corresponds to the standard model without demographic structure).

In summary we have determined the influence of recombination-rate fluctuations on the decorrelation of gene histories. We find that small-scale fluctuations are irrelevant to long-range correlation decay. Empirically determined large-scale fluctuations of the recombination rate, however, are found to significantly increase the correlations. Within a model with demographic structure, large-scale fluctuations of empirically determined recombination rates significantly contribute to the empirically observed slow decay of correlations.

We conclude by discussing the implications of our results for the study of genome-wide variability as reflected in single-nucleotide polymorphism (SNP) statistics. Eq. (1) determines the effect of recombination-rate fluctuations on $\rho^{\exp}_{\tau_x,\tau_y}$. This quantity, in turn, determines the genome-wide statistics of SNP locations: the variance of the number of SNPs in bins of lengths l along the chromosomes is determined by the integral of $\rho^{\exp}_{\tau_x,\tau_y}$ over x and y from 0 to l, i.e. by how fast the correlations decay on scales of length l (HUDSON, 1990). THE INTERNATIONAL SNP MAP WORKING GROUP (2001) has empirically determined the variance of the number of SNPs in short reads (of average length 500bp), the result was found to be consistent with the standard, unstructured gene-genealogical model assuming a constant recombination rate. This is consistent with our results (Fig. 2b): on scales of the order of 500bp, the recombination-rate fluctuations have little effect on the correlation function. We expect, however, that in order to understand the statistics of SNP counts in longer bins, it will be necessary to account for long-range recombination-rate fluctuations.

CHAKRAVARTI, A., BUETOW, K. H., ANTONARAKIS, S. E., WABER, P. G., BOEHM, C. D., and

KAZAZIAN, H. H., 1984 Nonuniform recombination within the human β-globin gene cluster. Am. J. Human Genetics **36**: 1239–1258.

ERIKSSON, A. and MEHLIG, B., 2004 unpublished .

EYRE-WALKER, A., GAUT, R. L., HILTON, H., FELDMAN, D. L., and GAUT, B. S., 1998 Investigation of the bottleneck leading to the domestication of maize. PNAS **95**: 4441–4446.

GOLDSTEIN, D. B., 2001 Islands of linkage disequilibrium. Nature Genetics **29**: 109–111.

GRIFFITHS, R. C., 1981 Neutral 2-locus multiple allele models with recombination. Theor. Pop. Biol. **19**: 169–186.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. **23**: 183–201.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology **7**: 1–44.

JEFFREYS, A. J., KAUPPI, L., and NEUMANN, R., 2001 Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. Nature Genetics **29**: 217–222.

KAPLAN, N. and HUDSON, R. R., 1985 The use of sample genealogies for studying a selectively neutral m-loci model with recombination. Theor. Pop. Biol. **28**: 382–396.

KONG, A., GUDBJARTSSON, D. F., SAINZ, J., JONSDOTTIR, G. M., GUDJONSSON, S. A., RICHARDSSON, B., SIGURDARDOTTIR, S., BARNARD, J., HALLBECK, B., MASSON, G., SHLIEN, A., PALSSON, S. T., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R., and STEFANSSON, K., 2002 A high-resolution recombination map of the human genome. Nature Genetics **31**: 241–247.

NORDBORG, M. and TAVARÉ, S., 2002 Linkage disequlibrium: what history has to tell us. TRENDS in Genetics **18**: 83–85.

REICH, D. E., SCHAFFNER, S. F., DALY, M. J., MCVEAN, G., MULLIKIN, J. C., HIGGINS, J. M., RICHTER, D. J., LANDER, E. S., and ALTSHULER, D., 2002 Human genome sequence variation and the influence of gene history, mutation and recombination. Nature Genetics **32**: 135–142.

STUMPF, M. P. H. and GOLDSTEIN, D. L., 2003 Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Current Biology **13**: 1–8.

TAVARÉ, S., 1984 Lines of descent and genealogical processes, and their application in population genetics models. Theor. Pop. Biol. **26**: 119–164.

TESHIMA, K. and TAJIMA, F., 2002 The effect of migration during the divergence. Theor. Pop. Biol. **62**: 81–95.

THE INTERNATIONAL SNP MAP WORKING GROUP, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**: 928–933.

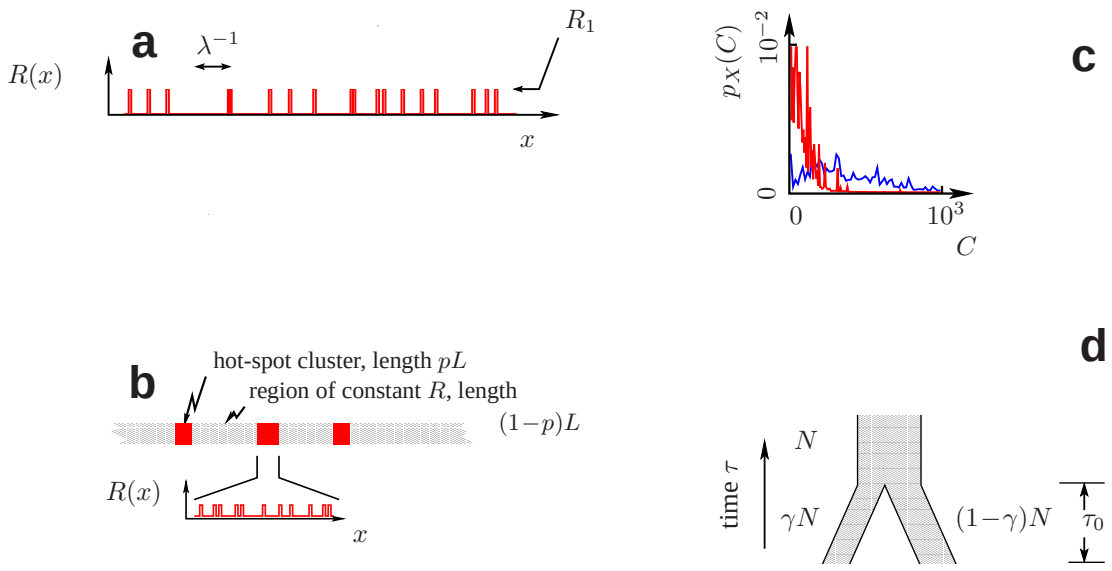

Figure 1: Description of models. **a** Model I: Recombination events occur at hot-spots (of zero width) with rate $R_1 = R/\lambda$. The remaining fraction occurs uniformly with rate $R_0 = (1-p)R$. The number of hot-spots in a locus of length l is Poisson distributed with rate λl. Eq. (1) gives $\rho^{\exp}_{\tau_x,\tau_y} = \mathrm{e}^{-\lambda X}\sum_{k=0}^{\infty}(\lambda X)^k/k!\,(R_1k+18)/[(R_1k)^2+13\,R_1k+18]$. This result is exact for $n = 2$. **b** Model II: hot-spots occur in clusters of size pL, separated by empty regions of length $(1-p)L$, $0 \le p \le 1$, and L is a typical length scale (of the order of several Mb). Within a cluster, the number of hot-spots is Poisson distributed. **c** Model III: the genome-wide distribution $p_X(C)$ is obtained by sampling $C = g(x+X) - g(x)$ from empirical data (KONG *et al.*, 2002) on the cumulative genetic distance $g(x)$ by randomly choosing physical positions x. The curve $g(x)$ is obtained from the Nature Genetics web supplement NG917-S13 (KONG *et al.*, 2002), from columns 1 (physical distance) and 3 (sex-averaged genetic distance) assuming an effective population size of $N = 10^4$, by ignoring entries labeled "NA", and by shifting the origin of both physical and genetic distances so that $g(0) = 0$. Shown here is $p_X(C)$ for chromosome 5; $X = 200$kb (red) and 1Mb (blue). **d** Demographic structure (divergent population): it is assumed that the population size was N until time τ_0 in the past when it split into two populations of sizes γN and $(1-\gamma)N$. The parameters τ_0 and N are chosen to be consistent with empirical data on the mean of the time to the most common recent ancestor and its coefficient of variation; table 1 in (REICH *et al.*, 2002). The asymmetry parameter varies between 0 and $1/2$. For sample size $n = 2$, the function $\rho_{\tau_x,\tau_y}(C)$ for this model was calculated by ERIKSSON and MEHLIG (2004).

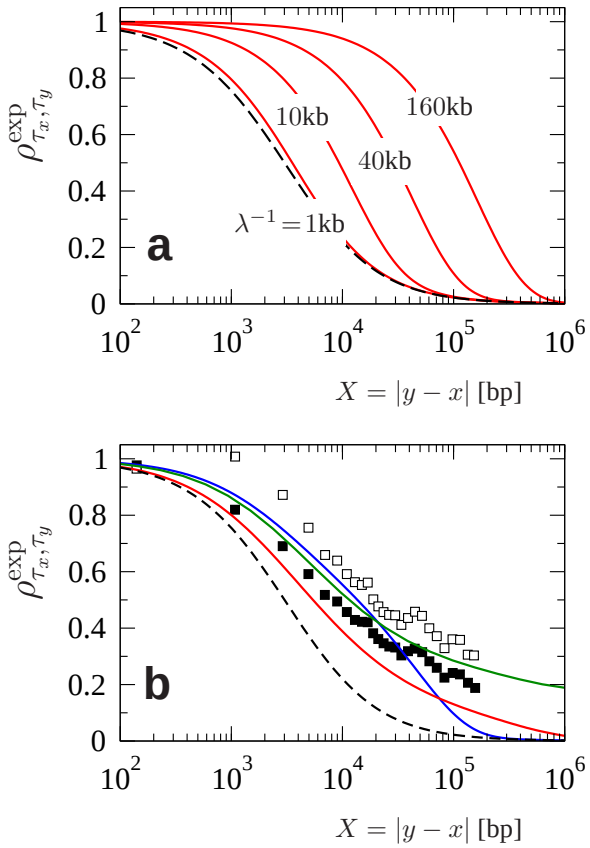

Figure 2: Decorrelation of gene histories $\rho^{\exp}_{\tau_x,\tau_y}$. **a** Model I (red lines), $R = 4Nr$, $N = 10^4$, and $r = 1.2$cM/Mb. The dashed line corresponds to constant recombination rate. **b** Model II (blue line), $p = 0.55, \lambda^{-1} = 50$kb, $r = 1.2$cM/Mb, and $L \gg X$, model III (red line), no fitting parameters, sex-averaged $p_X(C)$, model IV (green line) for $\gamma = 0.3$ (asymmetric split), empirical data (taken from Fig. 6a in (REICH *et al.*, 2002), upper and lower confidence limits, points), constant recombination rate (dashed line).